AD-A050 791 MASSACHUSETTS INST OF TECH CAMBRIDGE OPERATIONS RESE--ETC F/G 12/1
A SURVEY OF LAGRANGIAN TECHNIQUES FOR DISCRETE OPTIMIZATION.(U)
MAY 77 J F SHAPIRO DAAG29-76-C-0064
UNCLASSIFIED TR-133 ARO-14261.5-M NL

| OF |
AD
A050791

END
DATE
FILMED
4-78
DDC

AD No.______
DDC FILE COPY

AD A050791

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER 14 TR-133
Technical Report No. 133

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
6 A SURVEY OF LAGRANGEAN TECHNIQUES FOR DISCRETE OPTIMIZATION.

5. TYPE OF REPORT & PERIOD COVERED
9 Technical Report.
May 1977

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
10 Jeremy F. Shapiro

8. CONTRACT OR GRANT NUMBER(s)
15 DAAG29-76-C-0064

9. PERFORMING ORGANIZATION NAME AND ADDRESS
M.I.T. Operations Research Center ✓
77 Massachusetts Avenue
Cambridge, MA 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Office - Durham
Box CM, Duke Station
Durham, NC 27706

12. REPORT DATE
11 May 1977

13. NUMBER OF PAGES
41 pages 12 45P.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
18 ARO 19 14261.5-M

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Releasable without limitation on dissemination.

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

DDC
RECEIVED
MAR 6 1978
A

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Lagrangean Techniques
Integer Programming
Duality Theory
Branch-and-Bound
Subgradient Optimization
Group Theoretic Methods
Cutting Plane Method

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See page ii.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

270 720 JOB

A SURVEY OF LAGRANGEAN TECHNIQUES FOR DISCRETE OPTIMIZATION

by

JEREMY F. SHAPIRO

Technical Report No. 133

Work Performed Under
Contract DAAG29-76-C-0064, Army Research Office - Durham
"Basic Studies in Combinatorial and Nondifferentiable Optimization"
M.I.T. OSP 84475

Operations Research Center
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ACCESSION for
NTIS White Section ☒
DDC Buff Section □
UNANNOUNCED □
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

May 1977

Reproduction in whole or in part is permitted for any purpose of the United States Government.

FOREWORD

The Operations Research Center at the Massachusetts Institute of Technology is an interdepartmental activity devoted to graduate education and research in the field of operations research. The work of the Center is supported, in part, by government contracts and industrial grants-in-aid. The work reported herein was supported (in part) by the U.S. Army Research Office under Contract DAAG29-76-C-0064.

The author wishes to thank Paulo Villela for several important suggestions.

Jeremy F. Shapiro
Acting Director

ABSTRACT

This survey covers the theory and application of Lagrangean techniques to discrete optimization problems. A discussion of the applications includes integer programming special structures which can be exploited by Lagrangean techniques, multi-item production scheduling and inventory control problems, and the traveling salesman problem. The relationship of Lagrangean techniques to duality theory and convex analysis is given including a discussion of algorithms to solve the dual problems. Duality theory for integer programming and its relationship to the cutting plane method is reviewed. The use of Lagrangean techniques in conjunction with branch and bound is presented in a general framework for solving discrete optimization problems.

1. Introduction

Lagrangean techniques were proposed for discrete optimization problems as far back as 1955 when Lorie and Savage suggested a simple method for trying to solve zero-one integer programming (IP) problems. We use their method as a starting point for discussing many of the developments since then. The behavior of Lagrangean techniques in analyzing and solving zero-one IP problems is typical of their use on other discrete optimization problems discussed in later sections.

Specifically, consider the zero-one IP problem

$$\begin{aligned} v = \min\ & cx \\ \text{s.t. } & Ax \le b \\ & x_j = 0 \text{ or } 1. \end{aligned} \tag{1}$$

Let c_j denote a component of c, a_j a column of A with components a_{ij}, and b_i a component of b. Letting $\bar{u}$ represent a non-negative vector of Lagrange multipliers on the right hand side b, the method proceeds by computing the Lagrangean

$$L^o(\bar{u}) = -\bar{u}b + \underset{x_j = 0 \text{ or } 1}{\text{minimum}} \{(c+\bar{u}A)x\}. \tag{2}$$

The function $L^o(\bar{u})$ is clearly optimized by any zero-one solution $\bar{x}$ satisfying

$$\bar{x} = \begin{cases} 0 & \text{if } c_j+\bar{u}a_j > 0 \\ 0 \text{ or } 1 & \text{if } c_j+\bar{u}a_j = 0 \\ 1 & \text{if } c_j+\bar{u}a_j < 0 \end{cases} \tag{3}$$

In the introduction, we pose and discuss a number of questions about this method and its relevance to optimizing the original IP problem (1). In several instances, we will state results without proof. These results will either be proven in later sections, or reference will be given to papers containing the relevant proofs.

> *When is a zero-one solution $\bar{x}$ which is optimal in the Lagrangean also optimal in the given IP problem?*

In order to answer this question, we must recognize that the underlying goal of Lagrangean techniques is to try to establish the following sufficient optimality conditions.

OPTIMALITY CONDITIONS: The pair $(\bar{x},\bar{u})$, where $\bar{x}$ is zero-one and $\bar{u}\geq 0$, is said to satisfy the optimality conditions for the zero-one IP problem (1) if

$$\text{(i)} \quad L^{0}(\bar{u}) = -\bar{u}b + (c+\bar{u}A)\bar{x}$$

$$\text{(ii)} \quad \bar{u}(A\bar{x}-b) = 0$$

$$\text{(iii)} \quad A\bar{x} \leq b.$$

It can be shown that if the zero-one solution $\bar{x}$ satisfies the optimality conditions for some $\bar{u}$, then $\bar{x}$ is optimal in problem (1). This result is demonstrated in greater generality in section 3. The implication for the Lagrangean analysis is that $\bar{x}$ computed by (3) is optimal in problem (1) if it satisfies $A\bar{x} \leq b$ with equality on rows where $\bar{u}_i > 0$.

Of course, we should not expect that $\bar{x}$ computed by (3) will even be feasible in (1), much less optimal. According to the optimality conditions, however, such an $\bar{x}$ is <u>optimal</u> in any zero-one IP problem derived from (1) by replacing b with $A\bar{x} + \delta$ where δ is any non-negative vector satisfying $\delta_i = 0$ for i such that $\bar{u}_i > 0$. This property of $\bar{x}$ makes Lagrangean techniques useful in computing zero-one solutions to IP problems with soft constraints or in parametric analysis of an IP problem over a family of right hand sides. Parametric analysis of discrete optimization problems is discussed again in section 6.

> *How should the vector $\bar{u}$ of Lagrange multipliers be selected? Can we guarantee that there will be one which produces an optimal solution to the original IP problem?*

An arbitrary $\bar{u}$ fails to produce an optimal $\bar{x}$ because $\sum_j a_{ij}\bar{x}_j > b_i$ on some rows or $\sum_j a_{ij}\bar{x}_j < b_i$ on other rows with $\bar{u}_i > 0$. In order to change $\bar{u}$ so that the resulting $\bar{x}$ is closer to being feasible and/or optimal, we could consider increasing $\bar{u}_i$ on the former rows and decreasing $\bar{u}_i$ on the latter rows. A convergent tâtonnement approach of this type is non-trivial to construct because we must simultaneously deal with desired changes on a number of rows. Systematic adjustment of $\bar{u}$ can be achieved, however, by recognizing that there is a dual problem and a duality theory underlying the Lagrangean techniques. We discuss this point here briefly and in more detail in section 3.

For any $u \geq 0$, it can easily be shown that $L^o(u)$ is a lower bound on v, the minimal IP objective function cost in (1). The best choice of u is any one which yields the greatest lower bound, or equivalently, any u which is optimal in the dual problem

$$\begin{aligned} d^o &= \max L^o(u) \\ &\text{s.t. } u \geq 0. \end{aligned} \tag{4}$$

The reason for this choice is that if $\bar{u}$ can yield by (3) an optimal $\bar{x}$ to the primal problem (1), then $\bar{u}$ is optimal in (4). The validity of this statement can be verified by direct appeal to the optimality conditions using the weak duality condition $L^o(u) \leq v$ for any $u \geq 0$. Thus, a strategy for trying to solve the primal problem (1) is to compute an optimal solution $\bar{u}$ to the dual problem (4), and then try to find a complementary zero-one solution $\bar{x}$ for which the optimality conditions hold.

A fundamental question about Lagrangean techniques is whether or not an optimal dual solution to (4) can be guaranteed to produce an optimal solution to the primal IP problem (1). It turns out that the answer is no, although fail-safe methods exist and will be discussed for using the dual to solve the primal. If (4) cannot produce an optimal solution to (1), we say there is a duality gap.

Insight into why a duality gap occurs is gained by observing that problem (4) is equivalent to the LP dual of the LP relaxation of (1) which results by replacing $x_j = 0$ or 1 by the constraints $0 \leq x_j \leq 1$. This was first pointed out by Nemhauser and Ullman (1968). Here we use the term relaxation in the formal sense; that is, a mathematical programming problem is a relaxation of another given problem if its set of feasible solutions contains the set of feasible solutions to the given problem. The fact that dualization of (1) is equivalent to convexification of it is no accident because the equivalence of these two operations is valid for arbitrary mathematical programming problems (see Magnanti, Shapiro and Wagner (1976)). For discrete optimization problems, the convexified relaxations are LP problems. Geoffrion (1974) has used the expression Lagrangean relaxation to describe this equivalence. Insights and solution methods for the primal problem are derived from both the dualization and convexification viewpoints.

How should the dual problem be solved?

We remarked above that problem (4) is nothing more than the dual to the ordinary LP relaxation of (1). Thus, a vector of optimal dual variables could be calculated by applying the simplex algorithm to the LP relaxation of (1). The use of Lagrangean techniques as a distinct approach to discrete optimization has proven theoretically and computationally important for three reasons. First, dual problems derived from more complex discrete optimization problems than (1) can be represented as LP problems, but ones of immense size which cannot be explicitly constructed and then solved by the simplex method. Included in this category are dual problems more complex than (4) derived from (1) when (4) fails to solve (1). These are discussed in sections 2 and 4. From this point of view, the Lagrangean techniques applied to discrete optimization problems are a special case of dual decomposition methods for large scale LP problems (e.g., see Lasdon(1970)).

A second reason for considering the application of Lagrangean techniques to dual problems, in addition to the simplex method, is that the simplex method is exact and the dual problems are relaxation approximations. It is sometimes more effective to use an approximate method to compute quickly a good, but non-optimal, solution to a dual problem. In section 3, we consider alternative methods to the simplex method for solving dual problems and discuss their relation to the simplex method. The underlying idea is to treat dual problems as nondifferentiable steepest ascent problems taking into account the fact that the Lagrangean L^o is concave.

Lagrangean techniques as a distinct approach to discrete optimization problems emphasizes the need they satisfy to exploit special structures which arise in various models. This point is discussed in more detail in section 2.

What should be done if there is a duality gap?

As we shall see in section 3, a duality gap manifests itself by the computation of a fractional solution to the LP relaxation of problem (1). When this occurs, there are two complementary approaches which permit the analysis of problem (1) to continue and an optimal solution to be calculated. One approach is to branch on a variable at a fractional level in the LP relaxation; namely, use branch and bound. The integration of Lagrangean techniques with branch and bound is given in section 5. The other approach is to strengthen the dual problem (4) is by restricting the solutions permitted in the Lagrangean minimization to be a strict subset of the zero-one solutions. This is accomplished in a systematic fashion by the use of group theory and is discussed in section 4.

2. Exploiting Special Structures

Lagrangean techniques can be used to exploit special structures arising in IP and discrete optimization problems to construct efficient computational schemes. Moreover, identification and exploitation of special structures often provide insights into how discrete optimization models can be extended in new and richer applications.

The class of problems we consider first is

$$\begin{aligned} v = \min\ & cx \\ \text{s.t. } & Ax \leq b \\ & x \varepsilon X \subseteq R^n, \end{aligned} \tag{5}$$

where X is a discrete set with special structure. For example, X may consist of the multiple-choice constraints

$$\begin{aligned} \sum_{j \varepsilon J_k} x_j &= 1 \quad \text{for all } k \\ x_j &= 0 \text{ or } 1, \end{aligned} \tag{6}$$

where the sets J_k are disjoint. Another example occurs when X corresponds to a network optimization problem. In this case, the representation of X can either be as a totally unimodular system of linear inequalities , or as a network. Other IP examples are discussed by Geoffrion (1974).

The Lagrangean derived from (5) for any $u \geq 0$ is

$$L(u) = -ub + \min_{x \varepsilon X}(c+uA)x. \tag{7}$$

We expect L(u) to be much easier to compute than v because of the special form of X. Depending on the structure of X, the specific algorithm used to compute L may be a "good" algorithm in the terminology of Edmonds (1971) or Karp (1975); that is, the number of elementary operations required to compute L(u) is bounded by a polynomial of parameters of the problem. Even if it is not "good" in a strictly theoretical sense, the algorithm may be quite efficient empirically and derived from some simple dynamic programming recursion or list processing scheme. Examples will be given later in this section. Finally, in most instances the x calculations in (7) will be integer and may provide a useful starting point for heuristic methods to compute good solutions to (5).

Most discrete optimization problems stated in general terms can be formulated as IP problems, although sometimes with difficulty and inefficiently. We illustrate with two examples how Lagrangean techniques are useful in handling special structures which are poorly represented by systems of linear inequalities.

We consider a manufacturing system consisting of I items for which production is to be scheduled at minimum cost over T time periods. The demand for item i in period t is the nonnegative integer r_{it}; this demand must be met by stock from inventory or by production during the period. Let the variable x_{it} denote the production of item i in period t. The inventory of item i at the end of period t is

$$y_{it} = y_{i,t-1} + x_{it} - r_{it} \qquad t=1,\ldots,T$$

where we assume $y_{i,0} = 0$, or equivalently, initial inventory has been netted out of the r_{it}. Associated with x_{it} is a direct unit cost of production c_{it}. Similarly, associated with y_{it} is a direct unit cost of holding inventory h_{it}. The problem is complicated by the fact that positive *production of item i in period t* uses up a quantity $a_i + b_i x_{it}$ of a scarce resource q_t to be shared among the I items. The parameters a_i and b_i are assumed to be nonnegative. The use of Lagrangean techniques on this type of problem was originally proposed by Manne (1958). The model and analysis was extended by D. Zielinski and Gomory (1965) and has been applied by Lasdon and Terjung (1971).

This problem can be written as the mixed integer programming problem

$$v = \textit{minimum} \sum_{i=1}^{I} \sum_{t=1}^{T} (c_{it}x_{it}+h_{it}y_{it}) \tag{8a}$$

$$\text{s.t.} \sum_{i=1}^{I} (a_i\delta_{it}+b_ix_{it}) \le q_t\,, \quad t = 1,\ldots,T; \tag{8b}$$

for i = 1,...,I

$$\sum_{t=1}^{s} x_{it} - y_{is} = \sum_{t=1}^{s} r_{it} \qquad (8c)$$

$$s = 1,\ldots,T$$

$$x_{it} \le M_{it}\delta_{it} \; , \; t = 1,\ldots,T \qquad (8d)$$

$$x_{it} \ge 0 \; , \; y_{it} \ge 0 \qquad (8e)$$

$$\delta_{it} = 0 \text{ or } 1, \quad t = 1,\ldots,T$$

where $M_{it} = \sum_{s=t}^{T} r_{is}$ is an upper bound on the amount we would want to produce of i in period t. The constraints (8b) state that shared resource usage cannot exceed q_t. The constraints (8c) relate accumulated production and demand through period t to ending inventory in period t, and the nonnegativity of the y_{it} implies demand must be met and not delayed (backlogged). The constraints (8d) ensure that $\delta_{it} = 1$, and therefore the fixed charge resource usage a_i is incurred, if production x_{it} is positive in period t. Problem (8) is a mixed integer programming problem with IT zero-one variables, 2IT continuous variables and T + 2IT constraints. For the application of Lasdon and Terjung (1971), these figures are 240 zero-one variables, 480 continuous variables, and 486 constraints which is a mixed integer programming problem of significant size.

For future reference, define the set

$$N_i = \{(\delta_{it}, x_{it}, y_{it}) \; , \; t = 1,\ldots,T \,|\, \delta_{it}, x_{it}, y_{it} \text{ satisfy (8c), (8d), (8e)}\}. \qquad (9)$$

This set describes a feasible production schedule for item i ignoring the joint contraints (8b). The integer programming formulation (8) is not effective because it fails to exploit the special network structure of the sets N_i. This can be accomplished by Lagrangean techniques as follows. Assign Lagrange multipliers $u_t \geq 0$ to the scarce resources q_t and place the constraints (8b) in the objective function to form the Lagrangean

$$L(u) = - \sum_{t=1}^{T} u_t q_t$$

$$+ \underset{(\delta_{it}, x_{it}, y_{it}) \varepsilon N_i}{\text{minimum}} \sum_{i=1}^{I} \sum_{t=1}^{T} \{(c_{it} + u_t b_i) x_{it} + u_t a_i \delta_{it} + h_{it} y_{it}\}.$$

Letting

$$L_i(u) = \underset{(\delta_{it}, x_{it}, y_{it}) \varepsilon N_i}{\text{minimum}} \sum_{t=1}^{T} \{(c_{it} + u_t b_i) x_{it} + u_t a_i \delta_{it} + h_{it} y_{it}\}, \qquad (10)$$

the Lagrangean function clearly separates to become

$$L(u) = - \sum_{t=1}^{T} u_t q_t + \sum_{i=1}^{I} L_i(u).$$

Each of the problems (10) is a simple dynamic programming shortest-route calculation for scheduling item i where the Lagrange multipliers on shared resources adjust the costs as shown. Notice that it is easy to add any additional constraint on the problem of scheduling item i which can be accommodated by the network representation; for example, permitting production in period t only if inventory falls below a preassigned level.

Unfortunately, we must give up something in using Lagrangean techniques on the mixed IP (8) to exploit the special structure of the sets N_i. In the context of this application, the optimality conditions we seek but may not achieve involve Lagrange multipliers which permit each of the I items to be separately scheduled by the dynamic programming calculation L_i while achieving a global minimum. As we see in the next section, this can be at least approximately accomplished if the number of joint constraints is small relative to I.

In summary, the application of Lagrangean techniques just discussed involves the synthesis of a number of simple dynamic programming models under joint constraints into a more complex model. In a similar fashion, Fisher (1973) applied Lagrangean techniques to problems where a number of jobs are to be scheduled, each according to a precedence or CPM network, and the joint constraints are machine capacity. Another example is the cutting stock problem of Gilmore and Gomory (1963). In this model, a knapsack problem is used to generate cutting patterns and the joint constraints are on demand to be satisfied by some combination of the patterns generated.

The traveling salesman problem is a less obvious case where an underlying graph structure can be exploited to provide effective computational procedures. The problem is defined over a complete graph g with n nodes and symmetric lengths $c_{ij} = c_{ji}$ for all edges <i,j>. The objective is to find a minimum length tour of the n nodes, or in other words, a simple cycle of n edges and minimal length. This problem has several IP formulations involving $\frac{n(n-1)}{2}$ variables x_{ij} for the $n\frac{(n-1)}{2}$ edges <i,j> in the complete graph.

One such IP formulation consists of approximately 2^n constraints ensuring for feasible subgraphs of n edges that (i) the degree at each node is 2 and (ii) no cycle is formed among a subset of the nodes excluding node 1. The set of subgraphs of n edges satisfying (ii) has a very efficient characterization.

A <u>1-tree</u> defined on the graph g is a subgraph which is a tree on the nodes $2,\ldots,n$ and which is connected to node 1 by exactly two edges. The collection of subgraphs of n edges satisfying (ii) is precisely the set τ of 1-trees. Thus, the traveling salesman problem can be written as the IP problem

$$v = \min \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}x_{ij}$$

$$\text{s.t.} \quad \sum_{k<i} x_{ki} + \sum_{j>i} x_{ij} = 2 \qquad i=1,\ldots,n \tag{11}$$

$$x\varepsilon\tau \subseteq R^{\frac{n(n-1)}{2}} .$$

The implication of the formulation (11), however, is that we wish to deal with the 1-tree constraints implicitly rather than as a system of linear inequalities involving the zero-one variables x_{ij}.

Held and Karp (1970) discovered this partitioning of the traveling salesman problem and suggested the use of Lagrange multipliers on the degree constraints. For $u\varepsilon R^n$, the Lagrangean is

$$L(u) = -2\sum_{i=1}^{n} u_i + \underset{x\varepsilon\tau}{\text{minimum}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (c_{ij} + u_i + u_j)x_{ij} \tag{12}$$

This calculation is particularly easy to perform because it is essentially the problem of finding a minimum spanning tree in a graph. A "greedy" or "myopic" algorithm is available for this problem which is "good" in the theoretical sense and very efficient empirically (Kruskal(1956) and Edmonds(1971)).

The traveling salesman problem is only a substructure arising in applications of discrete optimization including vehicle routing and chemical reactor sequencing. Lagrangean techniques can be used to synthesize the routing or scheduling problems into a more complex model. In a similar application, Gomory and Hu(1964) discovered the importance of the spanning tree as a substructure arising in the synthesis of communications networks. Specifically, a maximum spanning tree problem is solved to determine the capacities in communications links required for the attainment of desired levels of flows. Lagrangean techniques can be used to iteratively select the spanning tree on which to perform the analysis until the communications problem is solved at minimum cost.

All of the special structures discussed above arise naturally in applications. By contrast, a recent approach to IP involves the construction of a special structure which we use as a surrogate for the constraints $Ax \leq b$. The approach requires that A and b in problem (1) have integer coefficients; henceforth we assume this to be the case. For expositional convenience, we rewrite the inequalities as equalities $Ax + Is = b$ where now we require the slack variables to be integer because A and b are integer.

The system $Ax + Is = b$ is aggregated to form a system of linear congruences which we view as an equation defined over a finite abelian group. The idea of using group theory to analyze IP problems was first suggested by Gomory (1965), although his specific approach was very different. We follow here the approach of Bell and Shapiro (1976). Specifically, consider the abelian group $G = Z_{q_1} \oplus Z_{q_2} \oplus \ldots \oplus Z_{q_r}$ where the q_i are integers greater than 1, Z_{q_i} is the cyclic group of order q_{i_1} and "$\oplus$" denotes direct sum. Let Z^m denote the set of all integer m-vectors, and construct a homomorphism ϕ from Z^m into G as follows. For each row i, we associate an element

ε of G and for any $z \varepsilon Z^m$, $\phi(z) = \sum_{i=1}^{m} z_i \varepsilon_i$. We apply ϕ to both sides of the linear system Ax +Is = b to aggregate it into the group equation $\sum_{j=1}^{n} \alpha_j x_j + \sum_{i=1}^{m} \varepsilon_i s_i = \beta$, where $\alpha_j = \phi(a_j)$, $\beta = \phi(b)$. It is easy to see that any integer x,s satisfying Ax + Is = b also satisfies the group equation. Therefore, we can add the group equation to the zero-one IP problem (1) without eliminating any feasible solutions. This gives us

$$\min cx \tag{13a}$$

$$\text{s.t.} \quad Ax + Is = b \tag{13b}$$

$$\sum_{j=1}^{n} \alpha_j x_j + \sum_{i=1}^{m} \varepsilon_i s_i = \beta \tag{13c}$$

$$x_j = 0 \text{ or } 1, \; s_i = 0,1,2,.... \tag{13d}$$

For future reference, let

$$\bar{Y} = \{(x,s) \mid (x,s) \text{ satisfies (13c) and (13d)}\}.$$

Lagrangean techniques are applied by dualizing with respect to the original constraints Ax + Is = b. For u≥0, this gives us the Lagrangean

$$L(u) = -ub + \underset{(x,s)\varepsilon\bar{Y}}{\text{minimum}} \{(c+uA)x + us\} \tag{14}$$

The calculation (14) can be carried out quite efficiently by list processing algorithms with the computation time determined mainly by the order of the group. (See Shapiro (1968), Glover (1969), Gorry, Northrup and Shapiro (1973).) It is easy to see that for a non-trivial group G (i.e., $|G| \geq 2$), the Lagrangean L gives higher lower bounds than L^o from section 1 since not all zero-one solutions x are included in $\bar{\underline{Y}}$ for some value of s. The selection of G and homomorphism ϕ is discussed again in section 4.

We have not attempted to be exhaustive in our discussion of the various discrete optimization models for which Lagrangean techniques have been successfully applied. Lagrangean techniques have also been applied to scheduling nuclear reactors (Mukstadt(1977)), the generalized assignment problem (Ross and Soland(1975)) and multi-commodity flow problems (Held, Wolfe and Crowder(1974)).

3. Duality Theory and the Calculation of Lagrange Multipliers

Implicit in the use of Lagrangean techniques is a duality theory for the optimal selection of the multipliers. We study this theory by considering the discrete optimization problem in the general form

$$\begin{aligned} v = \min\ & f(x) \\ \text{s.t. }\ & g(x) \leq b \\ & x \in X \subseteq R^n, \end{aligned} \tag{15}$$

where f is a scalar valued function defined on R^n, g is a function from R^n to R^m, and X is a discrete set. If there is no $x \in X$ satisfying $g(x) \leq b$, we take $v = +\infty$. With very little loss of generality, we assume that X is a finite set; say $X = \{x^t\}_{t=1}^{T}$. Implicit in this formulation is the assumption that the constraints $g(x) \leq b$ make the problem substantially more difficult to solve. Lagrangean techniques are applied by putting non-negative multipliers on the constraints $g(x) \leq b$ and placing them in the objective function.

Thus, the Lagrangean function derived from (15) is

$$\begin{aligned} L(u) &= -ub + \underset{x \epsilon X}{\text{minimum}} \{f(x) + ug(x)\} \\ &= -ub + \underset{t=1,\ldots,T}{\text{minimum}} \{f(x^t) + ug(x^t)\} \end{aligned} \tag{16}$$

As we saw in the previous section, the specific algorithm used to compute L may be a "good" algorithm, but even if it is not good, the intention is that it is quite efficient empirically and derived from a simple dynamic programming recursion or list processing scheme. Since X is finite, L is real valued for all u. Moreover, it is a continuous, but nondifferentiable, concave function (Rockafellar (1970)).

The combinatorial nature of the algorithms used in the Lagrangean calculation is a main distinguishing characteristic of the use of Lagrangean techniques in discrete optimization. This is in contrast to the application of Lagrangean techniques in nonlinear programming where f and g are differentiable, $X = R^n$ and the Lagrangean is minimized by solving the nonlinear system $\nabla f(x) + u\nabla g(x) = 0$. A second distinguishing characteristic of the use of Lagrangean techniques in discrete optimization is the non-differentiability of L, due to the discreteness of X. This makes the dual problem discussed below a non-differentiable optimization problem.

As it was for the zero-one IP problem discussed in the introduction, the selection of u in the Lagrangean is dictated by our desire to establish sufficient optimality conditions for (15).

OPTIMALITY CONDITIONS: The pair $(\bar{x},\bar{u})$, where $\bar{x} \epsilon X$ and $\bar{u} \geq 0$, is said to satisfy the optimality conditions for the discrete optimization problem (15) if

$$\text{(i)}\quad L(\bar{u}) = -\bar{u}b + f(\bar{x}) + \bar{u}g(\bar{x})$$

$$\text{(ii)}\quad \bar{u}(g(\bar{x}) - b) = 0$$

$$\text{(iii)}\quad g(\bar{x}) \le b.$$

Theorem 1: If $(\bar{x},\bar{u})$ satisfy the optimality conditions for the discrete optimization problem (15), then $\bar{x}$ is optimal in (15).

Proof: The solution $\bar{x}$ is clearly feasible in (15) since $\bar{x}\varepsilon X$ and $g(\bar{x}) \le b$ by condition (iii). Let $\tilde{x}\varepsilon X$ be any other feasible solution in (15). Then by condition (i),

$$L(\bar{u}) = -\bar{u}b + f(\bar{x}) + \bar{u}g(\bar{x}) \le -\bar{u}b + f(\tilde{x}) + \bar{u}g(\tilde{x}) \le f(\tilde{x}),$$

where the final inequality follows because $u\ge0$ and $g(\tilde{x}) - b\le0$. But by condition (ii), $L(\bar{u}) = f(\bar{x})$ and therefore $f(\bar{x}) \le f(\tilde{x})$ for all feasible $\tilde{x}$.||

Implicit in the proof of theorem 1 was a proof of the following important result.

Corollary 1 (weak duality). For any $u\ge0$, $L(u) \le v$.

Our primary goal in selecting u is to find one providing the greatest lower bound, or in other words, one which is optimal in the dual problem

$$d = \max_{s.t.\ u\ge0} L(u) \tag{17}$$

Corollary 2. If $(\bar{x},\bar{u})$ satisfy the optimality conditions for the discrete optimization problem (15), then $\bar{u}$ is optimal in the dual problem (17).

Proof: We have $L(\bar{u}) = -\bar{u}b + f(\bar{x}) + \bar{u}g(\bar{x}) = f(\bar{x}) = v$ by theorem 1. Since $L(u) \le v$ for all $u\ge0$ by corollary 1, we have $L(u) \le L(\bar{u})$ for all $u\ge0$.||

Thus, the indicated strategy for the application of Lagrangean techniques is to first find an optimal $\bar{u}$ in the dual problem. Once this has been done, we then try to find a complementary $\bar{x}\epsilon X$ for which the optimality conditions hold by calculating one or more solutions x satisfying $L(\bar{u}) = -\bar{u}b+f(x)+\bar{u}g(x)$. There is no guarantee that this strategy will succeed because (a) there may be no $\bar{u}$ optimal in the dual for which the optimality conditions can be made to hold for some $\bar{x}\epsilon X$; (b) the specific optimal $\bar{u}$ we calculated does not admit the optimality conditions for any $\bar{x}\epsilon X$; or (c) the specific $\bar{x}$ (or $\bar{x}$'s) in X selected by minimizing the Lagrangean do not satisfy the optimality conditions although some other $\tilde{x}\epsilon X$ which is minimal in the Lagrangean does satisfy them.

Lagrangean techniques can be applied in a fail-safe manner, however, if they are embedded in branch and bound searches. This is discussed in section 5. Alternatively, for some discrete optimization problems, it is possible to strengthen the dual problem if it fails to yield an optimal solution to the primal problem. Under certain conditions, the dual can be successively strengthened until the optimality conditions are found to hold. This is discussed in section 4.

For any $u \geq 0$, it is easy to see by direct appeal to the optimality conditions that $\bar{x}$ satisfying $L(\bar{u}) = -\bar{u}b+f(\bar{x}) + \bar{u}g(\bar{x})$ is optimal in (15) with b replaced by $g(\bar{x}) + \delta$ where δ is non-negative and satisfies $\delta_i = 0$ if $\bar{u}_i > 0$. Thus, Lagrangean techniques can be used in a heuristic manner to generate approximately optimal solutions to (15) when the constraints $g(x) \leq b$ are soft. Even if these constraints are not soft, heuristic methods exploiting the specific structure of (15) can be applied to perturb an infeasible $\bar{x}$ which almost satisfies the constraints to try to find a good feasible solution. D'Aversa(1977) has had success with this approach on IP problems.

There are two distinct but related approaches to solving (17); it can be viewed as a steepest ascent, nondifferentiable optimization problem, or as a large scale linear programming problem. We discuss first the steepest ascent approach. Our development is similar to that given in Fisher, Northup and Shapiro (1975); see also Grinold (1970), (1972) and Shapiro (1977). Although L is not everywhere differentiable, ascent methods can be constructed using a generalization of the gradient. An m-vector γ is called a subgradient of L at u if

$$L(u) \leq L(\bar{u}) + (u-\bar{u})\gamma \quad \text{for all } u.$$

For any subgradient, it can easily be shown that the half space $\{u|(u-\bar{u})\gamma \geq 0\}$ contains all solutions to the dual with higher values of L. In other words, any subgradient appears to point in a direction of ascent of L at $\bar{u}$. A readily available subgradient is

$$\bar{\gamma} = g(\bar{x}) - b \tag{18}$$

where $\bar{x}$ is any solution in X satisfying $L(\bar{u}) = -\bar{u}b+f(\bar{x})+\bar{u}g(\bar{x})$. If there is a unique $\bar{x}\epsilon X$ minimizing L at $\bar{u}$, then L is differentiable there and $\bar{\gamma}$ is the gradient.

The subgradient optimization method (Held and Karp (1971), Held, Wolfe and Crowder (1974)) uses these subgradients to generate a sequence $\{u^k\}$ of non-negative solutions to (17) by the rule

$$u_i^{k+1} = \max\ \{0, u_i^k + \theta_k \gamma_i^k\} \quad i=1,\ldots,m$$

where γ^k is any subgradient selected in (18) and $\theta_k > 0$ is the step length. For example, if θ_k obeys $\theta_k \rightarrow 0+$ and $\sum_k \theta_k \rightarrow +\infty$, then it can be shown that $L(u^k) \rightarrow d$ (Poljak (1967)). Alternatively, finite convergence to any target value $\bar{d}<d$ can be achieved if

$$\theta_k = \frac{\alpha_k(\bar{d}-L(u^k))}{||\gamma^k||^2} \tag{19}$$

where $||u^k||$ denotes Euclidean norm and $\varepsilon_1 < \alpha_k < 2-\varepsilon_2$ for $\varepsilon_1 > 0$, $\varepsilon_2 > 0$. The latter choice of θ_k is usually an uncertain calculation in practice, however, because the value d is not known and therefore a target value $\bar{d}<d$ cannot be chosen with certainty.

There is no guarantee when using subgradient optimization that $L(u^{k+1}) > L(u^k)$ although practice has shown that increasing lower bounds can be expected on most steps under the correct combination of artistic expertise and luck. Thus, subgradient optimization using the rule (19) for the step length is essentially a heuristic method with theoretical as well as empirical justification. It can be combined with convergent ascent methods for solving (17) based on the simplex method which we now discuss.

The dual problem (17) is equivalent to the LP problem

$$\begin{aligned} d &= \max \nu \\ \nu &\le -ub+f(x^t) + ug(x^t) \qquad t=1,\ldots,T \\ u &\ge 0, \end{aligned} \tag{20}$$

because, for any $u \ge 0$, the maximal choice $\nu(u)$ of ν is

$$-ub + \underset{t=1,\ldots,T}{\text{minimum}} \{f(x^t) + ug(x^t)\} = L(u).$$

Problem (20) is usually a large scale LP because the number T of constraints can easily be on the order of thousands or millions. For example, in the traveling salesman dual problem discussed in section 2, T equals the number of 1-trees defined on a graph of n nodes. The LP dual to (20) is

$$d = \min \sum_{t=1}^{T} f(x^t)\lambda_t \tag{21a}$$

$$\text{s.t.} \quad \sum_{t=1}^{T} g(x^t)\lambda_t \leq b \tag{21b}$$

$$\sum_{t=1}^{T} \lambda_t \quad = 1 \tag{21c}$$

$$\lambda_t \geq 0, \; t=1,\ldots,T \tag{21d}$$

This version of the dual problem clearly illustrates the nature of the convexification inherent in the application of Lagrangean techniques to the discrete optimization problem (15).

For decomposable discrete optimization problems with separable Lagrangeans such as the multi-item production scheduling and inventory control problem (10), the dual problem in the form (21) has a convexification constraint (21c) for each component in the Lagrangean. The number of such constraints for the production scheduling problem is I (the number of items being scheduled), and the number of joint constraints (21b) is T (one for each time period). If I > T, then an optimal solution to (21) found by a simplex algorithm will have pure strategies for at least I - T items; that is, one λ variable equal to one for these items. If I >> T, then Lagrangean techniques give a good approximation to an optimal solution to (10) because pure strategies are selected for most of the items. Roughly speaking, when I >> T, the duality gap between (10) and its dual is small.

Solution of the dual problem in its LP form (20) or (21) can be accomplished by a number of algorithms. One possibility is generalized linear programming, otherwise known as Dantzig-Wolfe decomposition (see Dantzig and Wolfe (1960), Lasdon (1970), or Magnanti, Shapiro and Wagner (1976)). Generalized linear programming proceeds by solving (21) with a subset of the T columns; this LP problem is called the Master Problem. A potential new column for the Master is generated by finding $\bar{x} \varepsilon X$ satisfying $L(-\bar{\pi}) = + \bar{\pi}b + f(\bar{x}) - \bar{\pi}g(\bar{x})$, where $\bar{\pi} \leq 0$ is the vector of optimal LP shadow prices on rows (21b) calculated for the Master problem. If $L(-\bar{\pi}) < \bar{\pi}b + \bar{\theta}$, where $\bar{\theta}$ is the optimal shadow price on the convexity row (21c), then the new column

$$\begin{pmatrix} f(\bar{x}) \\ g(\bar{x}) \\ 1 \end{pmatrix}$$

is added to the Master with new λ variable. If $L(-\bar{\pi}) = \bar{\pi}b + \bar{\theta}$ (the ">" case is not possible), then the optimal solution to the Master is optimal in the version (21) of the dual problem.

Note that if we required λ_t to be integer in version (21) of the dual problem, then (21) would be equivalent to the primal problem (15). Moreover, the dual solves the primal problem (15) if there is exactly one λ_t at a positive level in the optimal solution to (20); say, $\lambda_r = 1$, $\lambda_t = 0$, $t \neq r$. In that case, $x^r \varepsilon X$ is the optimal solution to the primal problem and we have found it by the use of Lagrangean techniques. Conversely, suppose more than one λ_t is at a positive level in the optimal solution to (21), say $\lambda_1 > 0, \ldots, \lambda_r > 0$, $\lambda_t = 0$, $t \geq r+1$. Then in all likelihood the solution $\sum_{t=1}^{r} \lambda_t x^t$ is not in X since X is discrete and the dual problem has failed to yield an optimal solution to the primal problem (15). Even if $y = \sum_{t=1}^{r} \lambda_t x^t$ is in X, there is no guarantee that y is optimal because optimality conditions (ii) and (iii) can fail to hold for y. In the next section we discuss how this difficulty can be overcome, at least in theory, and in section 5 we discuss the use of Lagrangean techniques in conjunction with branch and bound.

Generalized linear programming has some drawbacks as a technique for generating Lagrange multipliers in discrete optimization. It has not performed consistently (Orchard-Hays (1968)) although recent modifications of the approach such as BOXSTEP (Hogan, Marsten and Blankenship (1975)) have performed better. A second difficulty is that it does not produce monotonically increasing lower bounds to the primal objective function minimum. Monotonically increasing bounds are desirable for computational efficiency when Lagrangean techniques are used with branch and bound. A hybrid approach that is under investigation is to use subgradient optimization on the dual problem as an opening strategy and then switch to generalized linear programming when it slows down or performs erratically. The hope is that the generalized linear programming algorithm will then perform well because the first Master LP will have an effective set of columns generated by subgradient optimization with which to optimize.

An alternative convergent algorithm for the dual problem is an ascent method based on a generalized version of the primal-dual simplex algorithm. We present this method mainly because it provides insights into the theory of nondifferentiable optimization which is central to the selection of Lagrange multipliers for discrete optimization problems. Its computational effectiveness is uncertain although it has been imnlemented successfully for IP dual problems (see Fisher, Northup and Shapiro(1975) which also contains proofs of assertions).

The idea of the primal-dual ascent algorithm can best be developed by considering a difficulty which can occur in trying to find a direction of ascent at a point $\bar{u}$ with positive components where L is non-differentiable. The situation is depicted in figure 1. The vectors γ^1 and γ^2 are distinct subgradients of L at $\bar{u}$ and they both point into the half space containing points u such that $L(u) \geq L(\bar{u})$. Neither of these subgradients points in a direction of ascent of L; the directions of ascent are given by the shaded

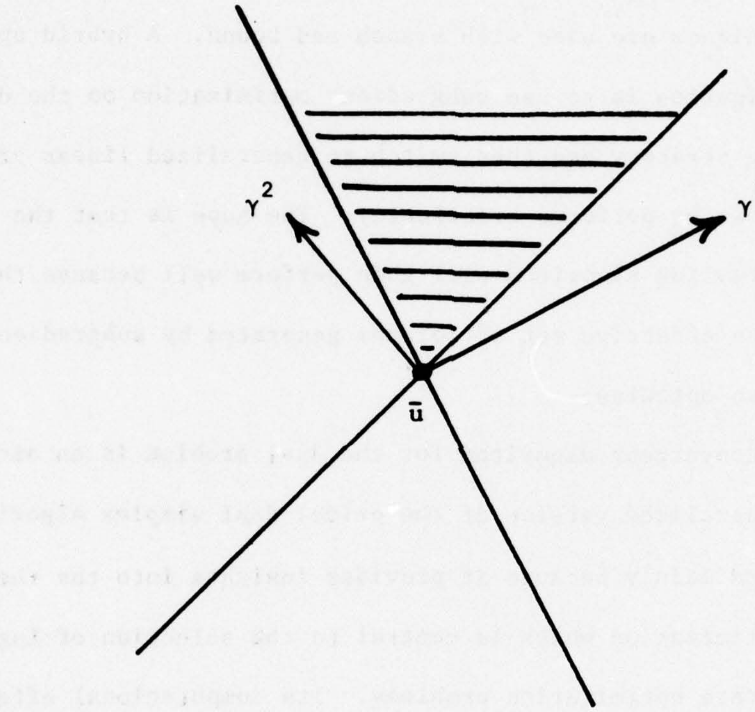

Figure 1

region which is the intersection of the two half spaces $\{u|(u-\bar{u})\gamma^1 \geq 0\}$ and $\{u|(u-\bar{u})\gamma^2 \geq 0\}$.

In general, feasible directions of ascent of L at a point $\bar{u}$ can be discovered only by considering at least implicitly the collection of all subgradients at that point. This set is called the <u>subdifferential</u> and denoted by $\delta L(\bar{u})$. The directional derivative $\nabla L(\bar{u};v)$ of L at the point $\bar{u}$ in the feasible direction v is given by (see Grinold(1970))

$$\nabla L(\bar{u};v) = \underset{\gamma\epsilon\delta L(\bar{u})}{\text{minimum}}\ v\gamma \tag{22}$$

The relation (22) is used to construct an LP problem yielding a feasible direction of ascent, if there is one; namely, a feasible direction v such that $\nabla L(\bar{u};v) > 0$. Two sets to be used in the construction of the direction finding LP are

$$V(\bar{u}) = \left\{ v\epsilon R^m \middle| \begin{array}{ll} 0 \leq v_i \leq 1 & \text{for i such that } \bar{u}_i = 0; \\ -1 \leq v_i \leq 1 & \text{for i such that } \bar{u}_i > 0 \end{array} \right\}$$

and

$$T(\bar{u}) = \{t|L(\bar{u}) = -\bar{u}b + f(x^t) + \bar{u}g(x^t)\}.$$

Without loss of generality, we can limit our search for a feasible direction of ascent to the set $V(\bar{u})$. The subdifferential $\delta L(\bar{u})$ is the convex hull of the points $\gamma^t = g(x^t) - b$, $t\epsilon T(\bar{u})$, and this permits us to characterize the directional derivative by the formula

$$\nabla L(\bar{u};v) = \underset{t\epsilon T(\bar{u})}{\text{minimum}}\ v\gamma^t \tag{23}$$

If the non-negative vector $\bar{u}$ is not optimal in the dual problem, then a direction of ascent of L at $\bar{u}$ can be found by solving the LP problem

$$\begin{aligned} \nabla &= \max \nu \\ \nu &\le v\gamma^t \ , \quad t\varepsilon T(\bar{u}) \\ &v\varepsilon V(\bar{u}). \end{aligned} \tag{24}$$

Conversely, if $\bar{u}$ is optimal in the dual problem, then $\nabla=0$ is the optimal objective function value in (24). Note that from (23) we have

$$\begin{aligned} \nabla &= \underset{v\varepsilon V(\bar{u})}{\text{maximum}} \ \underset{t=1,\ldots,T}{\text{minimum}} \ v\gamma^t \\ &= \underset{v\varepsilon V(\bar{u})}{\text{maximum}} \ \nabla L(\bar{u};v) \end{aligned}$$

and the LP (24) will pick out a direction of ascent with $\nabla L(\bar{u};v) > 0$ if there is one.

Once an ascent direction $v \neq 0$ with $\nabla L(\bar{u};v) = \nabla>0$ has been computed from (24), the step length $\theta>0$ in that direction is chosen to be the maximal value of θ satisfying $L(\bar{u}+\theta v) = L(\bar{u})+\theta\nabla$. This odd choice of θ is needed to ensure convergence of the ascent method by guaranteeing that the quantity ∇ strictly decreases from dual feasible point to dual feasible point (under the usual LP non-degeneracy assumptions). This is the criterion of the primal-dual simplex algorithm which in fact we are applying to the dual problem in the dual LP forms (20) and (21).

The difficulty with problem (24) is the possibly large number of constraints $\nu \le v\gamma^t$ since the set $T(\bar{u})$ can be large. This can be overcome by successive generation of rows for (24) as follows. Suppose we solve (24) with rows defined for γ^t, $t\varepsilon T'(\bar{u}) \subseteq T(\bar{u})$ and obtain an apparent direction v' of ascent satisfying $\nabla' = \underset{t\varepsilon T'(\bar{u})}{\text{minimum}} \ v'\gamma^t > 0$. We compute as before the maximal value θ' of θ such that $L(\bar{u}+\theta v') = L(\bar{u})+\theta\nabla'$. If $\theta' > 0$, then we proceed to $\bar{u} + \theta_{max} v$. If $\theta' = 0$, then it can be shown that we have found as a result of the line search a subgradient γ^s, $s \epsilon T(\bar{u}) - T'(\bar{u})$, which satisfies $\nabla' > v'\gamma^s$. In the latter case, we add $\nu \leq v\gamma^s$ to the abbreviated version of (24) and repeat the direction finding procedure by resolving it.

The dual problem is solved and $\bar{u}$ is found to be optimal if and only if $\nabla = 0$ in problem (24). Some additional insight is gained if we consider the LP dual to (24)

$$
\begin{aligned}
\nabla = \min \; & \sum_{i=1}^{m} s_i^- + \sum_{i \epsilon I^c(\bar{u})} s_i^+ \\
\text{s.t.} \; & \sum_{t \epsilon T(\bar{u})} \gamma_i^t \lambda_t - s_i^- + s_i^+ = 0 \qquad i = 1,\ldots,m \\
& \sum_{t \epsilon T(\bar{u})} \lambda_t = 1 \\
& \lambda_t \geq 0, \quad s_i^- \geq 0, \quad s_i^+ \geq 0,
\end{aligned}
\tag{25}
$$

where $I(\bar{u}) = \{i | \bar{u}_i = 0\}$. Problem (25) states, in effect, that $\bar{u}$ is optimal in the dual problem if and only if there exists a subgradient $\bar{\gamma} \epsilon \delta L(\bar{u})$ satisfying $\bar{\gamma}_i = 0$ for i such that $\bar{u}_i > 0$ and $\bar{\gamma}_i \leq 0$ for i such that $\bar{u}_i = 0$. Moreover, the columns γ^t for $t \epsilon T(\bar{u})$ and $\lambda_t > 0$ are an optimal set of columns for the dual problem in the form (21).

The close relationship among the concepts of dualization, convexification and the differentiability of L is again evident. Specifically, a sufficient, but not necessary, condition for there to be no duality gap is that L is differentiable at some optimal solution $\bar{u}$. If such is the case, then $\delta L(\bar{u})$ consists of the single vector $\bar{\gamma} = g(\bar{x}) - b$ which by necessity is the optimal column in (25). The necessary and sufficient condition for dual problem optimality that $\nabla = 0$ in (25) implies that $\bar{x}$ satisfies the optimality conditions implying it is optimal in the primal problem.

4. Resolution of Duality Gaps

We have mentioned several times in previous sections that a dual problem may fail to solve a given primal problem because of a duality gap. In this section, we examine how Lagrange techniques can be extended to guarantee the construction of a dual problem which yields an optimal solution to the primal problem. The development will be presented mainly in terms of the zero-one IP problem (1) as developed in Bell and Shapiro (1976), but the theory is applicable to the general class of discrete optimization problems (15). The theory permits the complete resolution of duality gaps. Nevertheless, numerical excesses can occur on some problems making it difficult in practice to follow the theory to its conclusion. The practical resolution of this difficulty is to imbed the dual analysis in branch and bound as described in section 5. Although it was not realized at the time it was invented, the cutting plane method for IP (Gomory (1958)) is a method for resolving IP duality gaps. We will make this connection in our development here, indicate why the cutting plane method proved to be empirically inefficient and argue that the dual approach to IP largely supercedes the cutting plane method.

We saw in section two how a dual problem to the zero-one IP problem (1) could be constructed using a group homomorphism to aggregate the equations $Ax \le b$. The relationship of this IP dual problem to problem (1) can be investigated using the duality theory developed in the previous section. Recall that the Lagrangean for the IP dual was defined for $u \ge 0$ as (see (14))

$$L(u) = -ub + \underset{(x,s)\varepsilon\bar{Y}}{\text{minimum}} \{(c+uA)x + us\},$$

where

$$\underline{\bar{Y}} = \{(x,s) \,|\, \sum_{j=1}^{n} \alpha_j x_j + \sum_{i=1}^{m} \varepsilon_i s_i = \beta, \;\; x_j\text{=0 or 1}, \; s_i\text{=0,1,2,....}\}. \qquad (26)$$

Although the slack variables in $\underline{\bar{Y}}$ are not explicitly bounded, we can without loss of generality limit $\underline{\bar{Y}}$ to a finite set, say $\underline{\bar{Y}} = \{(x^t,s^t)\}_{t=1}^{T}$. This is because any feasible slack vector $s = b-Ax$ is implicitly bounded by the zero-one constraints on x.

The general discrete optimization dual problem in the form (21) is specialized to give the following representation of the IP dual problem

$$\begin{aligned} d = \min \; & \sum_{t=1}^{T} (cx^t)\lambda_t \\ \text{s.t.} \; & \sum_{t=1}^{T} (Ax^t+Is^t)\lambda_t = b \\ & \sum_{t=1}^{T} \lambda_t \qquad\quad\;\; = 1 \\ & \qquad \lambda_t \geq 0 \end{aligned} \qquad (27)$$

This formulation of the IP dual problem provides us with the insights necessary to make several important connections between Lagrangean techniques and the cutting plane method. The convexification in problem (27) can be written in more compact form as

$$\begin{aligned} & d = \min cx \\ \text{s.t.} \;\; & x\varepsilon\{x|Ax+Is=b, \; 0\leq x_j\leq 1, \; 0\leq s_i\leq M_i\} \cap [\underline{\bar{Y}}], \end{aligned} \qquad (28)$$

where "[ ]" denotes convex hull and M_i is the upper bound on the slack variable s_i. In words, the IP dual problem is effectively the problem of minimizing cx over the intersection of the LP feasible region with the polyhedron $[\underline{\bar{Y}}]$. Inequalities based on the faces of $[\underline{\bar{Y}}]$ are cuts, and there will generally be an extremely large number of them. The computational inefficiency of the cutting plane method is due in large part to the algorithmic ambiguity created by this proliferation of cuts.

Lagrangean techniques and the IP dual problem provide a rationale for selecting cuts, but in the process, makes the use of cuts largely superfluous. For any $u \geq 0$, the inequality

$$(c+uA)x+us \geq L(u)+ub \qquad (29)$$

is a supporting hyperplane of $[\underline{\bar{Y}}]$. Since $\underline{\bar{Y}}$ contains all feasible solutions to the zero-one IP problem, (29) is a valid cut which can be added to any LP relaxation of the problem which included the constraints $Ax+Is = b$. Its effect on an LP relaxation would be to ensure that the objective function value cx would be at least $L(u)$ (Shapiro (1971)). Thus, the strongest cut in terms of forcing the objective function to increase is one derived from a dual vector u^* that is optimal in the dual problem. Furthermore, the procedure for selecting a cut according to this criterion is to solve the dual problem by one or more of the methods of the previous section which, as we see from problem (28), implicitly considers all cuts (i.e., all faces of $[\underline{\bar{Y}}]$) without generating any of them.

If an optimal solution to the dual problem produces an optimal solution to the zero-one IP problem, then a cut is not needed. If, on the other hand, an optimal zero-one solution is not produced, then a cut of the form (29) written with respect to an optimal dual solution u^* has the same effect on the objective function as all the cuts implied by $[\underline{\bar{Y}}]$. The addition of such a cut to an LP relaxation would permit the IP dual analysis to continue in the sense that a stronger IP dual of the form (27) could be derived. However, the construction of Bell and Shapiro (1977) attacks more directly the problem of strengthening the IP dual when it does not produce an optimal solution to the zero-one IP problem.

Solution of the zero-one IP problem (1) by Lagrangean techniques is constructively achieved by generating a finite sequence of groups $\{G^k\}_{k=0}^{K}$, sets $\{\bar{\underline{Y}}\}_{k=0}^{K}$ and IP dual problems analogous to (27) with objective function value d^k. The groups have the property that G^k is a subgroup of G^{k+1}, implying by the construction that $\bar{\underline{Y}}^{k+1} \subseteq \bar{\underline{Y}}^k$ and therefore that $v \geq d^{k+1} \geq d^k$. The critical step in this approach to solving (1) is that if an optimal solution to the kth dual does not yield an optimal solution to (1), then we are able to construct G^{k+1} so that $\bar{\underline{Y}}^{k+1} \subsetneq \bar{\underline{Y}}^k$. The construction uses as its point of departure the following result.

Theorem 2 (Bell and Shapiro): If only one λ_t is positive in an optimal basic solution to (27), then the corresponding solution (x^t,s^t) is optimal in the zero-one IP problem. On the other hand, if more than one λ_t is positive, then all the (x^t,s^t) corresponding to basic λ_t are infeasible in the zero-one IP problem.

When more than one λ_t is positive in an optimal basic solution to (27), then we can use a number theoretic procedure on the columns in (27) with λ_t positive to construct a new group with the property that the corresponding (x^t,s^t) are infeasible in the new group equation. Thus, they are not considered in the Lagrangean calculation. Since at least two solutions are eliminated each time the dual is strengthened, and since the set of (x,s) to be considered is finite, the process converges to an IP dual problem of the form (27) which yields an optimal solution to the zero-one IP problem.

Computational experience with the IP dual problem (27) is given in Fisher, Northup and Shapiro(1975). D'Aversa(1977) has encoded the iterative

IP dual analysis outlined above and experimentation is underway with it. The IP dual approach has been extended to mixed IP by Bell(1977) and Northup and Shapiro(1977). Burdet and Johnson(1975) have applied some concepts from convex analysis in the construction of IP methods which bear a resemblance to the methods just discussed.

The approach just outlined is applicable to the general discrete optimization problem (21) as long as $g(x^t)-b$ is a rational vector. If more than one λ_t is positive in (21), a group structure could be induced which would exclude infeasible solutions x^t from consideration in the Lagrangean (16). This would be accomplished by intersecting $\underline{X}$ with the set of solutions satisfying a group equation which would, however, make the algorithm for the Lagrangean more complex. See Bell(1973) for an application of this approach to the traveling salesman dual problem to maximize the Lagrangean (12).

5. Uses of Lagrangean techniques in branch and bound

Branch and bound is a method guaranteed to find an optimal solution to the general discrete optimization problem (15) by a systematic search of the discrete solution set X. The efficiency of the search is determined in large part by the strength of the bounds used in limiting it. Bounds are often derived from LP relaxations of a given discrete optimization problem which, as we have seen, arise naturally as dual problems for selecting Lagrange multipliers. Lagrangean analyses can also be used to indicate the most promising variables on which to branch. Conversely, branch and bound can be viewed as a method for perturbing a given discrete optimization problem when Lagrangean techniques fail to yield an optimal solution to it.

We describe the integration of Lagrangean techniques with branch and bound in terms of the general discrete optimization problem (15). Our development follows closely that of Fisher, Northup and Shapiro (1975). The branch and bound search of the set X is done in a non-redundant and implicitly exhaustive fashion. At any stage of computation, the least cost known solution $\hat{x} \varepsilon X$ satisfying $g(\hat{x}) \leq b$ is called the <u>incumbent</u> with <u>incumbent cost</u> $\hat{z} = f(\hat{x})$. Branch and bound generates a sequence of subproblems of the form

$$\begin{aligned} v(X^k) = \min\ & f(x) \\ \text{s.t. } & g(x) \leq b, \\ & x \varepsilon X^k, \end{aligned} \tag{30}$$

where $X^k \subseteq X$. The set X^k is selected to preserve the special structure of X. If we can find an optimal solution to (30), then we have implicitly tested all subproblems of the form (30) with X^k replaced by $X^\ell \subseteq X^k$ and such subproblems do not have to be explicitly enumerated. The same conclusion holds if we can ascertain that $v(X^k) \geq \hat{z}$ without actually discovering the precise value of $\overline{v}(X^k)$. If either of these two cases obtain, then we say that the subproblem (30) has been <u>fathomed</u>. If it is not fathomed, then we separate (30) into new subproblems of the form (30) with X^k replaced by X^ℓ, $\ell=1,\ldots,L$, and

$$\bigcup_{\ell=1}^{L} X^\ell = X^k, \quad X^{\ell_1} \cap X^{\ell_2} = \phi, \quad \ell_1 \neq \ell_2 .$$

Lagrangean techniques are used to try to fathom the subproblem (30) by solution of the dual problem

$$\begin{aligned} d(X^k) = \max\ & L(u;X^k) \\ \text{s.t. } & u \geq 0 , \end{aligned} \tag{31}$$

where

$$L(u;X^k) = -ub + \underset{x \varepsilon X^k}{\text{minimum}} \{f(x)+ug(x)\} . \tag{32}$$

The use of (31) in analyzing (30) is illustrated in figure 2 taken from Fisher, Northup and Shapiro (1975) which we now discuss step by step.

Steps 1 and 2: Often the inital subproblem list consists of only one subproblem corresponding to X.

Step 3: A good starting dual solution $\bar{u} \geq 0$ is usually available from previous computations.

Step 4: Computing the Lagrangean can be a network optimization problem, shortest route type computation for integer programming, minimum spanning tree for the traveling salesman problem, dynamic programming shortest route computation for resource constrained network scheduling problems, etc.

Step 5: As a result of step 4, the lower bound $L(\bar{u};X^k)$ on $v(X^k)$ is available, and it should be clear that (30) is fathomed if $L(\bar{u};X^k) \geq \hat{z}$ since $L(\bar{u};X^k) \leq v(X^k)$.

Steps 6, 7, 8: Let $\tilde{x} \varepsilon X^k$ be an optimal solution in (32) and suppose $\tilde{x}$ is feasible, i.e. $g(\tilde{x}) \leq b$. Since (30) was not fathomed (step 5), we have $L(\bar{u};X^k) = f(\tilde{x}) + \bar{u}(g(\tilde{x})-b) < \hat{z}$ with the quantity $\bar{u}(g(\tilde{x})-b) \leq 0$. Thus, it may or may not be true that $f(\tilde{x}) < \hat{z}$, but if so, then the incumbent $\hat{x}$ should be replaced by $\tilde{x}$. In any case, if $\tilde{x}$ is feasible, we have by the duality theory discussed in section 3 that $f(\tilde{x}) + \bar{u}(g(\tilde{x})-b) < v(X^k) \leq f(\tilde{x})$ and therefore $\tilde{x}$ is optimal in (30) if $\bar{u}(g(\tilde{x})-b) = 0$; i.e., if complementary slackness holds.

Step 9: This may be a test for optimality in the dual of the current $\bar{u}$. Alternatively, it may be a test of recent improvement in the

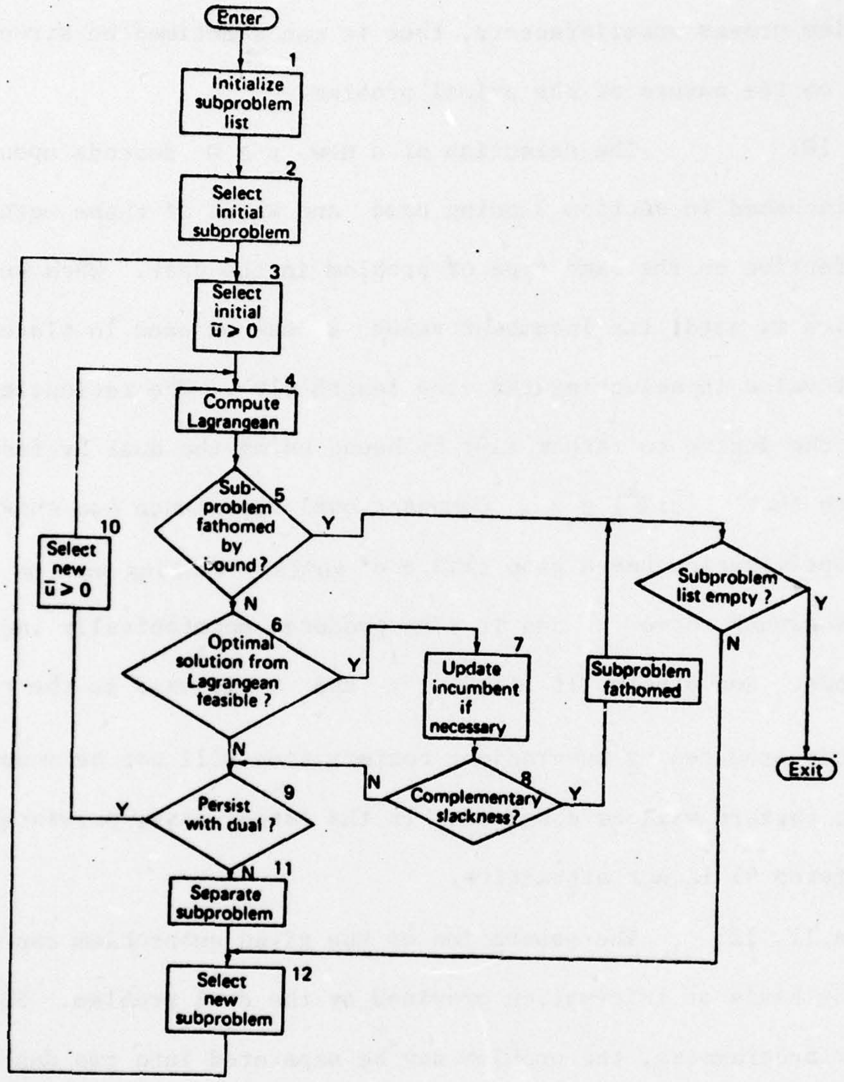
Enter
1
Initialize subproblem list
2
Select initial subproblem
3
Select initial ū > 0
4
Compute Lagrangean
5
Sub-problem fathomed by bound?
Y
N
10
Select new ū > 0
Subproblem list empty ?
Y
N
6
Optimal solution from Lagrangean feasible ?
Y
N
7
Update incumbent if necessary
Subproblem fathomed
Exit
8
Complementary slackness?
N
Y
9
Persist with dual ?
Y
N
11
Separate subproblem
12
Select new subproblem

Figure 2

dual lower bound. If generalized linear programming is used to solve the dual, then it provides at each iteration an upper bound $\bar{d}$ on $d(X^k)$. Thus, if $\bar{d} < \hat{z}$, we know that the subproblem (30) will never be fathomed by bound by the given dual. Finally, as we indicated in section 4, if the given dual problem proves unsatisfactory, then it can sometimes be strengthened depending on the nature of the primal problem.

Step 10: The selection of a new $\bar{u} \geq 0$ depends upon the methods discussed in section 3 being used and which of these methods have proven effective on the same type of problem in the past. When subgradient optimization is used, the incumbent value $\hat{z}$ can be used in place of $\bar{d}$ as the target value in selecting the step length (19). The rationale for this choice is the desire to fathom (30) by bound using the dual by finding $\bar{u} \geq 0$ such that $L(\bar{u};X^k) \geq \hat{z}$. Computational experience has shown that subgradient optimization has a good chance of quickly finding such a $\bar{u}$ if $d(X^k)$ is somewhat above $\hat{z}$ and it also produces monotonically increasing lower bounds. Conversely, if $d(X^k) < \hat{z}$ and $\hat{z}$ is used as the target, the lower bounds produced by subgradient optimization will not be monotonic and a wobbling pattern will be observed. In the latter case, persistence with the dual (step 9) is not attractive.

Steps 11, 12: The separation of the given subproblem can often be done on the basis of information provided by the dual problem. For example, in integer programming, the problem may be separated into two descendants with a zero-one variable x_j set to zero and one respectively, where x_j is chosen so that the reduced cost is minimal. It is important to point out that the greatest lower bound obtained during the dual analysis of (30) remains a valid lower bound on a subproblem derived from (30) with $X^\ell \subseteq X^k$.

In step 12, the new subproblem selected from the subproblem list can be one with low or minimal lower bound.

There are some constructs used in branch and bound derived from or related to Lagrangean techniques which we will not cover in any detail. One such construct is the calculation of penalties relative to a given LP relaxation of a discrete optimization problem (see Dakin(1965), Driebeek(1966), Healy(1964), Tomlin(1971)). A penalty for a zero-one IP problem, for example, is a lower bound estimate on the increase in cost of the primal objective function value as the result of separating the IP problem by fixing a specific variable at zero and one. Another construct is the surrogate constraint which is given in the form

$$f(x) + u(g(x)-b) < \hat{z}$$

for any $u \geq 0$ (see Geoffrion(1969) or Glover(1968)). The idea is that this constraint can be added to (30) since any feasible solution with lower cost than $\hat{z}$ will satisfy it. The constraint has a strong effect on the analysis of subproblems derived from (30) if u is chosen to be optimal or near optimal in the dual (31). Geoffrion(1974) discusses in greater detail penalties and surrogate constraints from the Lagrangean point of view.

6. Future research and applications areas

We have seen that Lagrangean techniques have already been widely used to analyze discrete optimization problems. Nevertheless, further progress should be possible in the use of these techniques, particularly in their integration with branch and bound, and the construction of fast hybrid algorithms for solving dual problems. We saw in section five that a family of related dual problems is generated and used in conjunction with branch and bound. The relationship between these duals is incompletely understood as are methods for exploiting the relationship in their optimization. Some work in this direction has been done by Marsten and Morin(1976). They give a new way to use linear programming to compute bounds on LP relaxations in branch and bound. Specifically, a resource-space tour is defined such that each simplex pivot yields a bound for every unfathomed subproblem in the branch and bound search.

Sensitivity and parametric analysis of IP problems is an area of current research interest and considerable practical importance in which Lagrangean techniques can play a significant role. Geoffrion and Nauss(1977) give an overview of the work done thus far in this area. Shapiro(1976) discusses how the constructs from section 4 can be used in sensitivity analysis. Multicriterion IP is a particularly desirable type of parametric analysis which has not yet been implemented. The idea would be to use the branch and bound search to generate a number of feasible IP solutions which are optimal or near optimal under various objective functions. The work required to find a number of interesting mixed IP solutions may be little more than that of finding a single optimal solution. Parametric variation of the right hand side is studied by Marsten and Morin(1975).

Another recent area of considerable research interest in which Lagrangean techniques are applicable is in the analysis of heuristic methods for combinatorial optimization. Cornuejols, Fisher and Nemhauser(1977) develop a "greedy" heuristic to generate feasible solutions to a class of location problems and use Lagrangean techniques to assess the error in objective function optimality.

References

[1] D.E. Bell (1973), "The resolution of duality gaps in discrete optimization", Tech. Report 81, M.I.T. Operations Research Center.

[2] D.E. Bell (1977), "Duality theory for MIP", (in preparation).

[3] D.E. Bell and J.F. Shapiro (1977), "A convergent duality theory for integer programming", to appear in Operations Research.

[4] C.A. Burdet and E.L. Johnson (1975), "A subadditive approach to solve linear integer programs", presented at Workshop on Integer Programming, Bonn.

[5] G. Cornuejols, M.L. Fisher and G.L. Nemhauser (1977), "Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms", Management Science 23, 789-810.

[6] R.J. Dakin (1965), "A tree search algorithm for mixed integer programming problems", Computer Journal 8, 250-255.

[7] G.B. Dantzig and P. Wolfe (1960), "Decomposition principle for linear programs", Operations Research 8, 101-111.

[8] J.S. D'Aversa (1977), "Computational experiments with IP duality", (Ph.D. thesis in preparation).

[9] N.J. Driebeek (1966), "An algorithm for the solution of mixed integer programming problems", Management Science 12, 576-587.

[10] B.P. Dzielinski and R. Gomory (1965),"Optimal programming of lot size inventory and labor allocations", Management Science 11, 874-890.

[11] J. Edmonds (1971), "Matroids and the Greedy algorithm", Mathematical Programming 1, 127-136.

[12] M.L. Fisher (1973), "Optimal solution of scheduling problems using Lagrange multipliers: Part I", Operations Research 21, 1114-1127.

[13] M.L. Fisher and J.F. Shapiro (1974), "Constructive duality in integer programming", SIAM Journal on Applied Mathematics 27, 31-52.

[14] M.L. Fisher, W.D. Northup and J.F. Shapiro (1975), "Using duality to solve discrete optimization problems: theory and computational experience", Mathematical Programming Study 3, 56-94.

[15] A.M. Geoffrion (1969), "An improved implicit enumeration approach for integer programming", Operations Research 17, 437-454.

[16] A.M. Geoffrion (1974), " Lagrangian relaxation and its uses in integer programming", Mathematical Programming Study 2, 82-114.

[17] A.M. Geoffrion and R. Nauss (1977), "Parametric and postoptimality analysis in integer linear programming", Management Science 23, 453-466.

[18] P.C. Gilmore and R.E. Gomory (1963), "A linear programming approach to the cutting-stock problem, Part II", Operations Research 11, 863-888.

References (continued)

[19] F. Glover (1968), "Surrogate constraints", Operations Research 16, 741-749.

[20] F. Glover (1969), "Integer programming over a finite additive group", SIAM Journal on Control 7, 213-231.

[21] R.E. Gomory (1958), "Essentials of an algorithm for integer solutions to linear programs", Bull. Amer. Math. Soc. 64, 275-278.

[22] R.E. Gomory and T.C. Hu (1964), "Synthesis of a communication network", SIAM Journal on Applied Mathematics 12, 348-389.

[23] R.E. Gomory (1965), "On the relation between integer and non-integer solutions to linear programs", Proc. Nat. Acad. Sci. 53, 260-265.

[24] G.A. Gorry, W.D. Northup and J.F. Shapiro (1973), "Computational experience with a group theoretic integer programming algorithm", Mathematical Programming 4, 171-192.

[25] R.C. Grinold (1970), "Lagrangean subgradients", Management Science 17, 185-188.

[26] R.C. Grinold (1972), "Steepest ascent for large scale linear programs", SIAM Review 14, 447-464.

[27] W.C. Healy, Jr. (1964), "Multiple choice programming", Operations Research 12, 122-138.

[28] M. Held and R.M. Karp (1970), "The traveling salesman problem and minimum spanning trees", Operations Research 18, 1138-1162.

[29] M. Held and R.M. Karp (1971), "The traveling salesman problem and minimum spanning trees: Part II", Mathematical Programming 1, 6-25.

[30] M. Held, P. Wolfe and H.D. Crowder (1974), "Validation of subgradient optimization", Mathematical Programming 6, 62-88.

[31] W.W. Hogan, R.E. Marsten and J.W. Blankenship (1975), "The Boxstep method for large scale optimization", Operations Research 23, (3).

[32] R.M. Karp (1975), "On the computational complexity of combinatorial problems", Networks 5, 45-68.

[33] J.B. Kruskal (1956), "On the shortest spanning subtree of a graph and the traveling salesman problem", Proc. Amer. Math. Soc. 2, 48-50.

[34] L.S. Lasdon (1970), Optimization theory for large systems (MacMillan).

[35] L.S. Lasdon and R.C. Terjung (1971), "An efficient algorithm for multi-item scheduling", Operations Research 19, 946-969.

[36] J. Lorie and L.I. Savage (1955), "Three problems in capital rationing", Journal of Business, 229-239.

[37] T.L. Magnanti, J.F. Shapiro and M.H. Wagner (1976), "Generalized linear programming solves the dual", Management Science 22, 1195-1204.

References (continued)

[38] A.S. Manne (1958), "Programming of economic lot sizes", Management Science 4, 115-135.

[39] R.E. Marsten and T.L. Morin (1975), "Parametric integer programming: the right-hand-side case", WP 808-75, Sloan School of Management, M.I.T.

[40] R.E. Marsten and T.L. Morin (1976), "A hybrid approach to discrete mathematical programming", OR 051-76, M.I.T. Operations Research Center.

[41] J. Mukstadt (1977), "An application of Lagrangian relaxation to scheduling in power generation systems", to appear in Operations Research.

[42] G.L. Nemhauser and Z. Ullman (1968), "A note on the generalized Lagrange multiplier solution to an integer programming problem", Operations Research 16, 450-452.

[43] W.D. Northup and J.F. Shapiro (1977), "A generalized linear programming algorithm for mixed integer programming", (in preparation).

[44] W. Orchard-Hays (1968), Advanced linear programming computing techniques (McGraw-Hill, New York).

[45] B.T. Poljak (1967), "A general method for solving extremum problems", Soviet Mathematics Doklady 8, 593-597.

[46] R.T. Rockafellar (1970), Convex analysis (Princeton University Press, Princeton, N.J.).

[47] G.T. Ross and R.M. Soland (1975), "A branch and bound algorithm for the generalized assignment problem", Mathematical Programming 8, 91-103.

[48] J.F. Shapiro (1968), "Dynamic programming algorithms for the integer programming problem I: the integer programming problem viewed as a knapsack type problem", Operations Research 16, 103-121.

[49] J.F. Shapiro (1971), "Generalized Lagrange multipliers in integer programming", Operations Research 19, 68-76.

[50] J.F. Shapiro (1976), "Sensitivity analysis in integer programming", OR 048-76, M.I.T. Operations Research Center.

[51] J.F. Shapiro (1977), Fundamental Structures of Mathematical Programming (text in preparation).

[52] J.A. Tomlin (1971), "An improved branch and bound method for integer programming", Operations Research 19, 1070-1075.